Journal of Research in Personality 47 (2013) 831–842

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

The inaccuracy of national character stereotypes

Robert R. McCrae [a], Wayne Chan [b], Lee Jussim [c], Filip De Fruyt [d], Corinna E. Löckenhoff [e], Marleen De Bolle [d], Paul T. Costa Jr. [b], Martina Hřebíčková [f], Sylvie Graf [f], Anu Realo [g], Jüri Allik [g,h], Katsuharu Nakazato [i], Yoshiko Shimonaka [j], Michelle Yik [k], Emília Ficková [l], Marina Brunner-Sciarra [m], Norma Reátigui [m], Nora Leibovich de Figueora [n], Vanina Schmidt [n], Chang-kyu Ahn [o], Hyun-nie Ahn [p], Maria E. Aguilar-Vafaie [q], Jerzy Siuta [r], Barbara Szmigielska [r], Thomas R. Cain [s], Jarret T. Crawford [t], Khairul Anwar Mastor [u], Jean-Pierre Rolland [v], Florence Nansubuga [w], Daniel R. Miramontez [x], Veronica Benet-Martínez [y], Jérôme Rossier [z], Denis Bratko [aa], Iris Marušić [ab], Jamin Halberstadt [ac], Mami Yamaguchi [ac], Goran Knežević [ad], Danka Purić [ad], Thomas A. Martin [ae], Mirona Gheorghiu [af], Peter B. Smith [ag], Claudio Barbaranelli [ah], Lei Wang [ai], Jane Shakespeare-Finch [aj], Margarida P. Lima [ak], Waldemar Klinkosz [al], Andrzej Sekowski [al], Lidia Alcalay [am], Franco Simonetti [am], Tatyana V. Avdeyeva [an], V.S. Pramila [ao], Antonio Terracciano [ap,b,*]

[a] 809 Evesham Avenue, Baltimore, MD, USA
[b] National Institute on Aging, Baltimore, MD, USA
[c] Department of Psychology, Rutgers University, New Brunswick, NJ, USA
[d] Department of Developmental, Personality, and Social Psychology, Ghent University, Ghent, Belgium
[e] Department of Human Development, Cornell University, Ithaca, NY, USA
[f] Institute of Psychology, Academy of Sciences of the Czech Republic, Brno, Czech Republic
[g] Department of Psychology, University of Tartu, Tartu, Estonia
[h] Estonian Academy of Sciences, Tallinn, Estonia
[i] Faculty of Social Welfare, Iwate Prefectural University, Iwate, Japan
[j] Department of Human Studies, Bunkyo Gakuin University, Bunkyo, Japan
[k] Division of Social Science, Hong Kong University of Science and Technology, Kowloon, Hong Kong
[l] Institute of Experimental Psychology, Slovak Academy of Sciences, Bratislava, Slovak Republic
[m] Faculty of Psychology, Universidad Peruana Cayetano Heredia, Lima, Peru
[n] Department of Psychology, University of Buenos Aires,Buenos Aires, Argentina
[o] Department of Education, Pusan National University, Busan, Republic of Korea
[p] Department of Psychology, Ewha Womans University, Seoul, Republic of Korea
[q] Department of Psychology, Tarbiat Modares University, Tehran, Iran
[r] Institute of Psychology, Jagiellonian University, Krakow, Poland
[s] School of Cognitive Science, Hampshire College, Amherst, MA, USA
[t] Department of Psychology, The College of New Jersey, Ewing, NJ, USA
[u] Personality Research Group, University Kebangsaan Malaysia, Bangi, Selangor, Malaysia
[v] UFR STAPS, Université Paris Ouest Nanterre La Défense, Nanterre, France
[w] Institute of Psychology, Makerere University, Kampala, Uganda
[x] Office of Institutional Research and Planning, San Diego Community College District, San Diego, CA, USA
[y] ICREA and Department of Political and Social Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[z] Institute of Psychology, University of Lausanne, Lausanne, Switzerland
[aa] Department of Psychology, University of Zagreb, Zagreb, Croatia
[ab] Institute for Social Research in Zagreb, Zagreb, Croatia
[ac] Department of Psychology, University of Otago, Dunedin, New Zealand
[ad] Department of Psychology, Belgrade University, Belgrade, Serbia
[ae] Department of Psychology, Susquehanna University, Selinsgrove, PA, USA
[af] School of Psychology, Queens University, Belfast, United Kingdom
[ag] School of Psychology, University of Sussex, Falmer, United Kingdom
[ah] Department of Psychology, La Sapienza University, Rome, Italy
[ai] Department of Psychology, Peking University, Beijing, China
[aj] School of Psychology and Counseling, Queensland University of Technology, Brisbane, Australia
[ak] Faculty of Psychology and Educational Science, University of Coimbra, Coimbra, Portugal
[al] Department of Psychology, John Paul II Catholic University of Lublin, Lublin, Poland
[am] Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile
[an] Graduate School of Professional Psychology, University of St. Thomas, Minneapolis, MN, USA

0092-6566/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jrp.2013.08.006

[ao] Visakhapatnam, Andhra Pradesh, India
[ap] Department of Geriatrics, Florida State University College of Medicine, Tallahassee, FL, USA

ARTICLE INFO

Article history:
Available online 22 August 2013

Keywords:
National character
Stereotypes
Five-Factor Model
Personality traits
Cross-cultural

ABSTRACT

Consensual stereotypes of some groups are relatively accurate, whereas others are not. Previous work suggesting that national character stereotypes are inaccurate has been criticized on several grounds. In this article we (a) provide arguments for the validity of assessed national mean trait levels as criteria for evaluating stereotype accuracy and (b) report new data on national character in 26 cultures from descriptions (N = 3323) of the typical male or female adolescent, adult, or old person in each. The average ratings were internally consistent and converged with independent stereotypes of the typical culture member, but were weakly related to objective assessments of personality. We argue that this conclusion is consistent with the broader literature on the inaccuracy of national character stereotypes.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Since Lippmann (1922/1991) first introduced the term *stereotype* to refer to people's beliefs about social groups, most social scientists have emphasized their inaccuracy (Allport, 1954/1979; Brown, 2010). Basic cognitive processes have been identified that lead people to exaggerate real differences between groups (Campbell, 1967), ignore or misremember stereotype-inconsistent information (Stangor & McMillan, 1992), and develop false beliefs to justify injustice (Jost & Banaji, 1994). These processes are practically important because of the role stereotypes can play in sustaining and exacerbating social inequalities, and theoretically important because they demonstrate that people's perceptions and judgments may deviate from objectivity and rationality.

"However", as Swim (1994, p. 21) put it, "reasons for inaccuracy are not evidence of inaccuracy". And, surprisingly, much of the evidence to date shows considerable accuracy in many consensual stereotypes, including those involving age (Chan et al., 2012), gender (Swim, 1994), and race (McCauley & Stitt, 1978; see reviews by Ryan, 2002; Jussim, 2012). By *accuracy* we mean statistical agreement between beliefs about a group and the aggregate characteristics of the group in question. Importantly, stereotype accuracy does not refer to beliefs about the sociological, historical, or biological bases of differences between groups; it implies only that individuals are able to perceive group differences with some degree of precision. We are concerned here with the accuracy of consensual stereotypes (operationalized as the average beliefs across a sample of respondents); because of the "wisdom of crowds" (Surowiecki, 2004) these are likely to be substantially more accurate than personal stereotypes.

It is now clear that the degree of accuracy or inaccuracy of stereotypes cannot be assumed, but must be evaluated empirically, on a case-by-case basis. In these evaluations, however, the burden of proof has shifted to those who claim that stereotypes are inaccurate, because failure to find evidence of accuracy is often a null result, and the interpretation of null results is always difficult. In this article we take on that burden with respect to the inaccuracy of national character stereotypes. We argue that aggregate personality traits are appropriate criteria for evaluating the accuracy of national character stereotypes and review evidence on the adequacy of our stereotype measure; we then report new data replicating previous findings of inaccuracy.

The term *national character* might be broadly understood to include a wide range of characteristics, including intelligence, appearance, food preferences, and athletic abilities (e.g., Ibrahim et al., 2010). We adopt a narrower view, equating character with personality traits, and we use a comprehensive model of personality traits, the Five-Factor Model (FFM). Our study thus speaks to the accuracy of national stereotypes of personality traits, but does not imply accuracy or inaccuracy in perceptions of other national characteristics.

Age and gender stereotypes concerning personality traits appear to be largely accurate (Chan et al., 2012; Löckenhoff et al., 2013), but Terracciano et al. (2005) reported that national character stereotypes are not. They examined beliefs about the typical personality traits of members of different cultures and found that they were essentially unrelated to assessed mean levels of traits in 49 cultures. However, that conclusion has been challenged on a number of grounds (e.g., Perugini & Richetin, 2007). Because stereotypes in general are often accurate, it is reasonable to ask if flaws in the Terracciano study accounted for the negative results. Because the sample was large (N = 3,989) and a number of alternative analytic strategies were employed, the most plausible arguments are that (a) the criteria—i.e., assessed national levels of personality traits—were invalid, or (b) the stereotype measure was inadequate. We consider these arguments and then offer new data on the (in)accuracy of national character stereotypes.

1.1. Validity of the accuracy criteria

The accuracy of stereotypes can only be determined by comparing beliefs to some objective standard. Terracciano et al. (2005) argued that objective data on the mean levels of personality traits in various nations provided such a standard, but that view is currently a matter of controversy. In the 2005 study, personality was assessed using either self-reports or observer ratings of individuals in each culture on versions of the NEO Inventories (McCrae & Costa, 2010), which measure 30 specific traits, or facets, that define the five major personality factors of the FFM. There is ample evidence that these instruments provide valid assessments of personality within cultures—that is, when members of a culture are compared to each other (e.g., McCrae & Terracciano, 2005a).

It is less certain that mean values can be compared across cultures, because different translations, response styles, or reference group effects (RGEs) may limit the scalar equivalence of scores (Church et al., 2011; Heine, Buchtel, & Norenzayan, 2008; Zecca et al., 2013). However, there are reasons to doubt that response styles or problems in translation have serious effects on culture-level NEO Inventory scores, as a number of studies have shown. We review these before turning to a consideration of the RGE.

There are known cultural differences in acquiescent responding (Smith, 2004), but scales from the NEO Inventories have balanced

* Corresponding author. Address: Department of Geriatrics, Florida State University College of Medicine, 1115 W. Call Street, Room 4425-B, Tallahassee, FL 32306-4300, USA.
E-mail address: antonio.terracciano@med.fsu.edu (A. Terracciano).

keying, so acquiescent responding should have minimal effect. Cultures also differ in self-enhancement; that might bias self-report data, but should not affect informant ratings of personality. Mõttus et al. (2012a) showed that extreme responding, although it had little effect on individual scores, had a larger effect on culture-level scores of Conscientiousness. Nevertheless, the rank-order of cultures was similar when scores corrected for extreme responding were compared to uncorrected scores, rho = .68, $p < .001$. The frequency of random responding or missing data might vary across cultures, but McCrae and colleagues (McCrae & Terracciano, 2005b; McCrae et al., 2010) screened out protocols with evidence of random responding or excessive missing data.

Several studies have asked bilingual respondents to complete inventories twice, in different languages. Using the Big Five Inventory (BFI; Benet-Martínez & John, 1998), Ramírez-Esparza, Gosling, Benet-Martínez, Potter, and Pennebaker (2006) showed that English/Spanish bilinguals scored higher on Extraversion, Agreeableness, and Conscientiousness when tested in English. However, most studies using the NEO Inventories have seen only small and scattered differences. For example, in a study of Hong Kong Chinese, consistent differences were found for only 3 of 30 facets (Excitement Seeking, Straightforwardness, and Altruism; McCrae, Yik, Trapnell, Bond, & Paulhus, 1998). Bilingual studies in Korean (Piedmont & Chae, 1997), Shona (Piedmont, Bain, McCrae, & Costa, 2002), and Spanish (Costa, McCrae, & Kay, 1995) have also shown comparability of mean levels across translations for most scales. Different Filipino samples completing the NEO Inventory in Filipino or in English (Church & Katigbak, 2002) and different Indian samples completing the inventory in Marathi or Telugu (McCrae, 2002) showed similar, although not identical, profiles.

These studies suggest that response styles and translations may have some impact on culture-level scores, but that it is likely to be relatively small. When culture-level scores are examined directly for construct validity—a "top down" approach—several lines of evidence support their validity. The geographical distribution of traits is consistent with the hypothesis that national scores are accurate reflections of trait levels (Allik & McCrae, 2004; see also Gelade, 2013)—for example, Danes and Norwegians showed similar personality profiles, as did Zimbabweans and Black South Africans. Again, scores are meaningfully correlated at the culture level with dimensions of culture; for example, cultures high in aggregate Openness score high in Hofstede's Individualism (Hofstede & McCrae, 2004).

Perhaps most compelling are data from three different sources that demonstrate mutual agreement. McCrae (2002) compiled self-report data from 36 cultures; McCrae and colleagues (McCrae & Terracciano, 2005b) obtained observer ratings of college-age and adult targets in 51 cultures as part of the Personality Profiles of Cultures (PPOC) project; and McCrae et al. (2010) gathered observer ratings of 12–17-year-old targets in 24 cultures as part of the Adolescent PPOC (APPOC). Correlations across cultures of mean scores from these three studies for each of the 30 facet scales of the NEO Inventories ranged from −.18 to .82 (*Mdn* = .52); 69 (77%) of these were statistically significant (McCrae & Terracciano, 2005b; McCrae et al., 2010). Additional analyses using intraclass correlations (*ICC*s) showed significant agreement for personality profiles within most cultures (68%). Schmitt et al. (2007) found evidence of convergent validity for Extraversion, Conscientiousness, and Neuroticism when BFI culture means were correlated with NEO Inventory means—although the scales showed rather poor discriminant validity in that study. McCrae (2002) found evidence for both convergent and discriminant validity when NEO Inventory means were correlated with Eysenck Personality Questionnaire (Eysenck & Eysenck, 1975) Neuroticism and Extraversion means (see also Bartram, 2013). Taken together, these findings appear to provide evidence of construct validity for the national means.[1]

1.2. The reference-group effect and other standards of comparison

However, Heine and colleagues (Heine, Lehman, Peng, & Greenholtz, 2002; Heine et al., 2008; Heine & Buchtel, 2009) have argued that these findings may be artifacts of the RGE. In this view, responses to personality items are not absolute judgments, but are made relative to some implicit normative group, notably the citizens of one's country: "Japanese tend to evaluate themselves on the basis of how they compare with other Japanese, whereas Canadians tend to evaluate themselves on the basis of how they compare to other Canadians" (Heine et al., 2002, p. 905). RGEs have been demonstrated to operate in several contexts; for example, Guimond et al. (2007) offered evidence that women in more traditional cultures describe themselves relative to other women, whereas women in more progressive cultures adopt people-in-general as their frame of reference.

Heine and colleagues further argued that RGEs can explain much of the "top down" evidence for the validity of culture means in personality traits. Data from self-reports agree with data from peer ratings, but this might be because both adopt the same reference group. Similarly, geographically close nations might have similar personality profiles because they share similar RGEs.

This is an appealing argument, but it requires careful scrutiny. A number of considerations argue against it.

(1) The first and most obvious problem is that, carried to its logical conclusion, RGE would eliminate cultural differences in assessed traits, because means everywhere would be average. About half the population in any culture would call themselves *high* on a trait (relative to their compatriots), and half would call themselves *low*; the culture mean would always be average. (Note that this is exactly what would happen if a researcher standardized raw scores as *T*-scores within each culture: All means would be 50.) Where there is no variation—except random sampling error—there can be no correlation, so we would not, for example, expect NEO Neuroticism means to be correlated across cultures with EPQ Neuroticism means, or with mean peer ratings of Neuroticism. Yet such correlations are repeatedly found.
Heine et al. (2002) recognized this problem and argued that consistent cultural differences occur because social comparison "is not the only process by which people come to understand themselves" (p. 907). This suggests that RGE serves only to attenuate cultural differences—to drive scores some way toward the mean. That would imply that assessed culture means are not accurate in an absolute sense, but—other things being equal—would still be accurate in a relative sense, and it is the relative levels of traits across cultures that we use to assess the accuracy of national character stereotypes.

(2) It is not clear to whom people in fact compare themselves when completing personality inventories. To their circle of friends? To the national average? To a "perceived international norm" (p. 301) that Heine et al. (2008) suggested is used when describing national stereotypes? In some cultures, comparisons seem to be made to one's own gender

[1] These studies do not directly address the issue of measurement invariance as assessed by such techniques as multigroup confirmatory factor analysis (MCFA). A head-to-head comparison of results from bottom-up techniques like MCFA with top-down approaches such as culture-level correlations would be illuminating.

(Guimond et al., 2007). The most plausible case is that different respondents choose different reference groups, contributing noise, but not systematic bias, to mean scores.

(3) RGE is more problematic for some items than for others. As Heine et al. (2002) noted, responses to some questions "might rely more on introspection and comparison with internal standards than on implicit comparisons with consensually shared standards" (p. 914). Rating the item "I am not a very methodical person" may require some idea of how methodical the typical person is, but it is not clear that any reference group at all is needed to rate such items as "I have never literally jumped for joy" or "I'd rather vacation at a popular beach than an isolated cabin in the woods." All these items are included in the NEO Inventories.

(4) The RGE is in fact only one example of a broader class of artifacts, namely, different standards of comparison. In particular, members of a culture might have very high standards for assessing a trait such as Conscientiousness, not because their compatriots on average scored high on the trait (as the RGE assumes), but because Conscientiousness is highly valued in the culture. But are there in fact large cultural differences in norms for Conscientiousness? Mõttus and colleagues (Mõttus et al., 2012b) examined that idea by generating a set of anchoring vignettes and asking respondents in 21 diverse cultures to rate the Conscientiousness of the individual depicted in each. They concluded that there were "no substantial culture-related differences in standards for Conscientiousness" (p. 303) and the small differences they found had little effect on the ranking of self-report means in these cultures.

(5) McCrae et al. (1998) examined social judgment effects by comparing ratings of Chinese undergraduates made by Canadian-born Chinese or recent immigrants from Hong Kong. These two groups of raters might be assumed to have different standards for judging personality traits. However, the resulting profiles were strikingly similar, and showed significant differences for only four of the 30 NEO Inventory facets and only one factor, Neuroticism (Hong Kong-born raters perceived Chinese undergraduates as somewhat higher in Neuroticism than did Canadian-born raters).

(6) Geographical patterns cannot easily be explained by RGEs or other culture differences in standards—at least not in ways that favor the accuracy of national stereotypes. Heine and colleagues argued that geographically close countries such as the US and Canada have similar observed mean personality profiles (*ICC* = .66; Terracciano et al., 2005) because they share cultural norms for the assessment of traits. Shared standards would indeed lead to similar observed profiles—but only if the real underlying profiles were also similar. This must be so because the observed score is a function of the true score and the standard of evaluation that is implicitly relied on in evaluating the true score. But the national stereotypes of unassuming Canadians and arrogant Americans are diametrically opposed (*ICC* = –.53); if the true underlying profiles are similar, one or both of the stereotypes must be wrong.

These arguments are not definitive. There have been no studies to date on cultural standards for four of the five factors. England and Australia might have different true score profiles *and* different standards of comparison that just happen to cancel out to yield similar observed profiles. But the most parsimonious conclusion at present is that RGE and other cultural differences in standards for evaluating traits have fairly minor effects. The assessed personality profiles in our criterion sets are surely not perfect, but they are probably adequate for the assessment of the accuracy of national stereotypes.

1.3. Adequacy of the stereotype measure

Is it possible that the null results reported by Terracciano et al. (2005) are due to problems in the instrument used to assess stereotypes, the National Character Survey (NCS)? When first used it was an ad hoc measure with 30 items corresponding conceptually to the 30 facet scales of the NEO Inventories. Respondents were asked to "judge the likelihood of 30 characteristics for the typical" member of their own culture, using five-point scales. For example, the characteristic national level of anxiety was rated on a scale from *anxious, nervous, worrying* to *at ease, calm, relaxed*. Terracciano et al. (2005) showed that the NCS had reasonable psychometric properties, given its brevity: The five domain scales (created by summing the relevant six items for each of the five factors) had adequate internal consistency, the factor structure gave a reasonable approximation to that of the NEO Inventories, and, when aggregated across raters, the mean scores for each trait reliably distinguished among nations.

Subsequent studies using the NCS have provided additional support. Terracciano and McCrae (2007) reported that NCS ratings of Americans remained similar in Lebanon (*ICC* = .74) and Italy (*ICC* = .92) in the six-month period before and after the American invasion of Iraq. Five years after the PPOC study, the NCS was readministered to new samples of raters in Estonia and Poland (Realo et al., 2009) and in Slovakia, Germany, Poland, and the Czech Republic (Hřebíčková & Graf, 2013; Kouřilová & Hřebíčková, 2011); in all five cultures, very similar trait profiles were found on the two occasions (*ICC*s = .78–.93, $N = 30$, $p < .001$). This might be seen as evidence of retest reliability at the culture level; it is a particularly stringent test, both because different raters were used on different occasions, and because the retest interval was quite long.

These two studies also provided evidence that different translations of the NCS yield comparable scores. Raters in Estonia, Finland, Belarus, Lithuania, Latvia, and Poland generally agreed on their depiction of the typical Russian (*Mdn ICC* = .58; Realo et al., 2009). Raters in Austria, Germany, the Czech Republic, Poland, and Slovakia generally agreed on their views of each other (e.g., Czechs and Germans agreed on the depiction of Austrians), with 25 of 30 comparisons statistically significant (*Mdn ICC* = .68; Hřebíčková & Graf, 2013). Further, there is evidence that heterostereotypes agree with autostereotypes for some (though not all) cultures (Boster & Maltseva, 2006). For example, pooled international ratings of the typical American closely resembled ratings from Americans themselves (Terracciano & McCrae, 2007), and the stereotype of Germans held by other central Europeans matched German autostereotypes (Hřebíčková & Graf, 2013). These findings might be interpreted as evidence of the interrater reliability of the NCS at the culture level, an international consensus on national stereotypes. But consensus is not necessarily evidence of accuracy (Kenny, 1994), just as reliability is not equivalent to validity.

It is more difficult to assess the validity of the NCS. A stereotype measure that accurately reflects what people believe might be called *valid*, even if the beliefs themselves were entirely false. To avoid confusion between validity and accuracy, we will refer to this psychometric property as *fidelity*: Does the NCS faithfully reflect the beliefs of respondents? On its face, it does. Terracciano et al. (2005) found, for example, that Americans were characterized as being assertive and the British were described as reserved; these seem to fit familiar stereotypes. Chinese Malaysian students stereotype Malays as friendly but lazy (Ibrahim et al., 2010), consistent with their NCS scores on Warmth ($T = 54.5$) and Self-Discipline ($T = 45.1$; McCrae, Terracciano, Realo, & Allik, 2007). Church and Katigbak (2002) recruited panels of American and Filipino judges, all of whom had lived in both the US and the

Philippines for at least three years, and asked them to indicate on a 7-point scale whether Filipinos or Americans were higher on each of the NEO facet traits. These judgments correlated $r = .72$, $N = 30$, $p < .001$, with the difference between Terracciano et al.' NCS scores for Filipinos and Americans. However, a broader and more systematic assessment of fidelity is needed.

When the validity of a trait measure is assessed, the most common form of evidence is a correlation between the measure and another scale designed to assess the same trait—for example, a new anxiety scale may be correlated with an established measure of anxiety. McCrae, Terracciano, Realo, and Allik (2008) provided such evidence for NCS measures of Agreeableness and Conscientiousness by correlating them with scales from the Global Leadership and Organizational Behavior Effectiveness (GLOBE) project (House, Hanges, Javidan, Dorfman, & Gupta, 2004). The GLOBE Humane Orientation scale asks informants if members of their culture are generous and friendly; this scale correlated .50 ($N = 33$ cultures, $p < .01$) with NCS Agreeableness. The GLOBE Future Orientation scale, which assesses the degree to which typical culture members are thought to plan ahead, correlated .65 with NCS Conscientiousness. At present, there do not appear to be alternative measures of national stereotypes of Neuroticism, Extraversion, or Openness (but see Peabody, 1985, for data on other personality variables). One design might be to ask respondents to complete the full, 240-item NEO Inventory to describe the typical culture member and to evaluate the briefer NCS against that criterion. In the present study we assess the fidelity of the NCS across different formats of administration. Our design allows us to ask if NCS scores faithfully reflect the perceived character of the whole nation, or if they depict only some demographic segments of the nation.

1.4. The accuracy of national character stereotypes

Terracciano et al. (2005) assessed the accuracy of national character autostereotypes—the views of the group held by ingroup members—of 30 traits across 49 cultures, and of 49 cultures across a 30-trait profile, using both self-reported and observer rated personality assessments as criteria. They found no consistent evidence of accuracy, except for the personality profile of Poles. McCrae et al. (2010) found that the national autostereotype profiles reported by Terracciano and colleagues were related to mean national profiles of adolescents aged 12–17 in Argentina ($ICC = .39$, $p < .05$) and Turkey ($ICC = .42$, $p < .05$), but not in 20 other cultures. Realo et al. (2009) reported the accuracy of autostereotypes of national character in nine samples from seven cultures, using the NCS to assess stereotypes and a modification of the NCS to obtain self-reported personality assessments to serve as the criteria. They showed agreement (ICCs = .39–.52) for only four of the nine samples (Poles, Finns, Russians, and adult Estonians). Hřebíčková and Graf (2013) found accurate autostereotypes for Poles and adult Czechs, but not for Austrians, Germans, Slovaks, or college-age Czechs. Overall, it appears that there is little evidence that national character autostereotypes as assessed by the NCS are accurate representations of mean trait levels except in Poland, where agreement may be simply a coincidence.

Arguably, autostereotypes may be distorted by ethnocentric bias, whereas the perceptions of outgroup members may be more objective. Terracciano et al. (2005) did not address that possibility, but Realo et al. (2009) compared perceptions of the typical Russian by Belarusians, Estonians, Finns, Latvians, Lithuanians, and Poles to the assessed personality profile of Russians and found no agreement (ICCs = −.39 to .31). Hřebíčková and Graf (2013) used the NCS to gather information on the views of Austrians, Czechs, Germans, Poles, and Slovaks about each other. They found no evidence that any of these heterostereotypes agreed with assessed personality. Outgroup members do not appear to have any more accurate perceptions of national character than do ingroup members.

Many stereotypes are reasonably accurate (Jussim, 2012). When the 30 NCS items were used to assess typical adolescents, adults, and old persons, these age stereotypes proved to be remarkably accurate when compared to known age differences in personality (Chan et al., 2012). In addition, Löckenhoff et al. (2013) showed that, when applied to males and females, NCS items captured gender stereotypes that correspond to established sex differences in personality (Costa, Terracciano, & McCrae, 2001). These studies, which also used NEO Inventory data as criteria, demonstrate that stereotypes assessed by the NCS may be quite accurate when an appropriate target is chosen. Of course, there is no guarantee that all stereotypes are accurate, and if NCS ratings of national character do not resemble NEO Inventory profiles of different cultures, it is probably because national character stereotypes are inherently inaccurate.

One possible explanation for that inaccuracy might be that national character stereotypes vary substantially across different subcultures or subgroups. For example, the stereotype of Northern Italians is dramatically different from that of Southern Italians (McCrae et al., 2007). Conceivably, national character may be different for males and females, or for adults and old persons. If, when asked to rate the typical culture member, some respondents use men as their frame of reference and others use women, the pooled responses might be meaningless.

1.5. A replication and extension

Psychologists have recently been reminded of the crucial importance of replication for their science (Pashler & Wagenmakers, 2012). It is thus important to attempt to replicate the null findings of Terracciano et al. (2005); here we use a modified version of the NCS in a subset of the cultures originally examined. As part of the APPOC (De Fruyt, De Bolle, McCrae, Terracciano, & Costa, 2009), respondents in 26 cultures were asked to make ratings of the typical male or female of a specific age in their own culture—for example, one group of Ugandans rated the traits of the typical adolescent Ugandan girl. Earlier research had asked only about an undifferentiated national character (e.g., the typical Ugandan), and it is unclear what respondents had in mind when making ratings. In the present study age and gender of the target are specified, and we can determine if national stereotypes are in fact consistent across these categories. If national stereotypes prove to be generalizable across age and gender categories, then averaging them may provide the most faithful assessment of the true national stereotype. The accuracy of these assessments can be judged against assessed mean personality traits. It has been shown that national personality profiles are generalizable across age and gender (McCrae, 2002; McCrae & Terracciano, 2005b), so the criteria can be averaged across these groups.

However, the personality profiles of different age and gender groups in a given culture are not identical, and it is possible that age- and gender-specific stereotypes will be more accurate when compared to criteria matched on age and gender. This hypothesis is based on the premise that people have extensive experience with different age and gender groups within their own culture, and can therefore describe age and sex differences with some degree of accuracy. When they assess a particular category (e.g., adult male Chileans), their ratings are a function of their accurate knowledge of age and gender differences as well as their beliefs about national differences. Even if their national character stereotypes are completely unfounded, these ratings will correlate to some extent with the assessed personality traits of the corresponding age and gender group, because both sets of scores share variance due to

true within-culture differences in age and gender. In this study, we also test that hypothesis.

2. Method

2.1. Procedure

Participants ($N = 3323$) from 26 countries around the world rated the personality characteristics of typical males and females in their culture as part of a study on stereotype accuracy (see Chan et al., 2012). These participants were previously described in detail in Löckenhoff et al. (2009); about two-thirds were women and most were in their early 20 s.[2] They were assigned either males or females as targets, then rated the personality traits of the typical adolescent, adult, and old person in their own country in counterbalanced order. Previous analyses (Chan et al., 2012) suggested that one effect of this design was to exaggerate age contrasts for the second and third rating; for that reason this article employs only the first rating made by each participant.[3] Approximately one-sixth of the sample provided ratings of each group: adolescent boys, adolescent girls, adult men, adult women, old men, and old women.[4]

In previous studies, Chan et al. (2012) and Löckenhoff et al. (2013) analyzed the accuracy of age and gender stereotypes, respectively, using the data set analyzed here. The present article provides the first analyses of the accuracy of national character stereotypes.

2.2. Measures

Stereotypes were assessed using the National Character Survey (NCS; Terracciano et al., 2005), which consists of 30 bipolar items corresponding to the 30 facets of the NEO Inventories (McCrae & Costa, 2010). NCS domain scores are calculated by summing six facets for each of the five factors; in this sample, Cronbach's αs ($N = 3323$) were .62, .61, .66, .66, .77 for Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, respectively, which are acceptable for six-item scales. Interjudge reliability [$ICC(1,k)$] of the mean values for each Culture $\times$ Sex $\times$ Age group was calculated for each NCS domain and facet; these values ranged from .64 to .95 ($Mdn = .80$), suggesting that respondents generally agree on the perceived personality features that differentiate these groups. Analyses also showed that male and female raters gave similar ratings (see Chan et al., 2012, for details).

To assess overall national character stereotypes, we collapsed data across age and gender groups. NCS scores were first standardized as *T*-scores ($M = 50$, $SD = 10$) within the entire sample, so that all traits would have the same metric. These scores, however, contain variance due to stereotypes of age and gender that can obscure the effects of national character. To eliminate universal age and gender effects we centered each trait within each of the six target groups by subtracting its mean within that target group and adding 50.0, such that the grand mean of each trait across all raters is 50.0 for each age and gender group. Note that this is similar to the familiar practice of standardizing personality test scores by using age- and gender-specific norms. The recentered NCS scores are age- and gender-corrected ratings that should contain only variance due to national character differences, age and gender effects unique to different cultures, and error. Finally, the recentered scores within rater country were averaged across all six targets to generate mean national character stereotype scores for each trait.

2.3. Comparison samples

To assess the *fidelity* of mean NCS scores as measures of national character stereotypes, we compared them to previously published mean national character stereotype data ($N = 25$ cultures with complete data; Terracciano et al., 2005). To assess the *accuracy* of national character stereotypes, we compared mean NCS scores to mean personality assessments from three previous studies that had used versions of the NEO Inventories. The first (Self) consisted of self-reports from college-age and adult respondents ($N = 18$ overlapping cultures; McCrae, 2002; McCrae & Terracciano, 2008). These data had been collected by a variety of independent investigators using their own sampling designs; some included only adult, some only student samples, and some both age groups. McCrae (2002) standardized the data as *T*-scores within age and gender groups using American norms (Costa & McCrae, 1992) and estimated national trait levels as the unweighted average of the available groups. The second comparison sample (PPOC) consisted of observer ratings of college-age and adult (40 + years) targets ($N = 26$ cultures; McCrae & Terracciano, 2005b). These data had been collected as part of a collaborative project in which roughly equal samples of college age and adult male and female targets had been rated. Data were standardized as *T*-scores using norms from the full international PPOC sample, and culture means were published in McCrae and Terracciano (2008). The third comparison sample (APPOC) consisted of observer ratings of adolescents aged 12–17 ($N = 21$ cultures; De Fruyt et al., 2009) collected in a design modeled on the PPOC. *T*-scores were calculated using the full APPOC norms (McCrae et al., 2010).

3. Results

3.1. Generalizability across targets

If national character stereotypes in a given culture are truly national, they ought to be reflected in perceptions of the typical man as well as the typical woman, the typical adolescent as well as the typical adult in that culture. To test that assumption, we conducted reliability analyses at the domain and profile levels, asking whether facet levels across all cultures were similar in each of the six target groups. For the five domain analyses, we treated recentered NCS facet means in each country as cases ($N = 6$ facets $\times$ 26 cultures = 156 rows) and the age-by-gender target groups as items ($k = 6$ columns). The composite score, which is the mean of the six groups (items), had an α of .65 for Neuroticism, suggesting that in cultures where adolescents were thought to be high in Neuroticism facets, adults were also generally thought to be high; where women were considered low in Neuroticism (relative to women in other cultures), so, in general, were men. Alphas for the other domains were .56, .71, .66, and .46 for Extraversion, Openness, Agreeableness, and Conscientiousness, respectively. For the full profile analysis, we treated recentered NCS facet means in each country as cases ($N = 30$ facets $\times$ 26 cultures = 780 rows) and the age-by-gender target groups as items ($k = 6$ columns). For this analysis, α was .62, which is adequate for so brief a measure. All corrected item/total correlations were above .25 ($p < .001$), and none of the six category means would improve α if removed from the scale. It thus appears to be appropriate to treat the mean across categories as a measure of overall national character stereotypes.

3.2. Fidelity and accuracy of stereotypes for traits

We analyzed fidelity and accuracy of scales on a case-by-case basis. For the trait analyses, the fidelity or accuracy of each domain and facet scale was calculated as the Pearson correlation between

[2] Supplementary analyses in a subset of cultures (Chan et al., 2012; Terracciano et al., 2005) suggested that similar results would be obtained from older raters.

[3] Analyses using all three ratings showed results similar to, but somewhat weaker than, those presented here.

[4] In France, all participants first rated the typical adolescent French boy or girl, so only these two groups are represented in the French composite.

Table 1
Correlations across cultures of APPOC NCS means with PPOC NCS and NEO Inventory means.

Trait	NCS	NEO Inventory		
	PPOC[a] (25)	Self[b] (18)	PPOC[c] (26)	APPOC[d] (21)
N_{NCS}: Neuroticism	.59***	.21	.39*	.30
E_{NCS}: Extraversion	.51**	.08	.19	−.23
O_{NCS}: Openness	.48**	−.42	−.10	−.04
A_{NCS}: Agreeableness	.67***	.39	.33*	.18
C_{NCS}: Conscientiousness	.45*	.11	.17	.45*
$N1_{NCS}$: Anxiety	.46**	−.17	−.02	.08
$N2_{NCS}$: Angry Hostility	.67***	.14	.45*	.39*
$N3_{NCS}$: Depression	.73***	−.09	.15	.05
$N4_{NCS}$: Self-consciousness	.31	.11	.11	.11
$N5_{NCS}$: Impulsiveness	.49**	−.06	.14	.05
$N6_{NCS}$: Vulnerability	.49**	.52*	.59***	.46*
$E1_{NCS}$: Warmth	.65***	−.16	−.41	−.45
$E2_{NCS}$: Gregariousness	.35*	.16	.09	.00
$E3_{NCS}$: Assertiveness	.24	−.24	−.23	−.21
$E4_{NCS}$: Activity	.28	−.26	.02	−.11
$E5_{NCS}$: Excitement Seeking	.45*	.05	.35*	.16
$E6_{NCS}$: Positive Emotions	.52**	−.09	.24	−.46
$O1_{NCS}$: Fantasy	.47**	−.08	.02	−.16
$O2_{NCS}$: Aesthetics	.47**	−.17	.30	.58**
$O3_{NCS}$: Feelings	.36*	−.60	−.50	−.49
$O4_{NCS}$: Actions	.41*	−.18	−.03	−.07
$O5_{NCS}$: Ideas	.44*	.08	.49**	.10
$O6_{NCS}$: Values	.42*	−.47	−.11	−.02
$A1_{NCS}$: Trust	.56**	.21	.23	−.13
$A2_{NCS}$: Straightforwardness	.59***	−.12	.10	.07
$A3_{NCS}$: Altruism	.39*	.14	.08	−.04
$A4_{NCS}$: Compliance	.56**	.18	.23	.25
$A5_{NCS}$: Modesty	.68***	.32	−.09	.12
$A6_{NCS}$: Tender-Mindedness	.65***	.62**	.65***	.25
$C1_{NCS}$: Competence	.19	.20	.01	.40*
$C2_{NCS}$: Order	.34*	.26	.42*	.40*
$C3_{NCS}$: Dutifulness	.35*	.04	.18	.44*
$C4_{NCS}$: Achievement Striving	.18	.04	.13	.14
$C5_{NCS}$: Self-Discipline	.63**	.07	.07	.29
$C6_{NCS}$: Deliberation	.64**	.27	.37*	.56**
Facet *Mdn*	.47	.05	.12	.09

Note: *Ns* in parentheses. APPOC = Adolescent Personality Profiles of Cultures project. PPOC = Personality Profiles of Cultures project. NCS = National Character Survey.
[a] Ratings of typical culture member from Terracciano et al. (2005).
[b] Self-reports from McCrae (2002), and McCrae and Terracciano (2008).
[c] Peer ratings from McCrae and Terracciano (2005b).
[d] Peer ratings from De Fruyt et al. (2009).
* $p < .05$, one-tailed.
** $p < .01$, one-tailed.
*** $p < .001$, one-tailed.

mean NCS scores from the present study and mean scores from the comparison samples (e.g., stereotypes of national Extraversion vs. mean assessments of Extraversion). This analysis allows us to determine whether national character stereotypes are faithful and accurate for specific personality traits. Note that these correlations do not depend on the metric of the scales (e.g., raw scores, American norms, international norms) because correlations are invariant across linear transformations.

Table 1 contains the comparison analyses by domain and facet. The first data column demonstrates convergent fidelity for stereotypes of specific traits. All correlations are positive, and 30 of 35 are statistically significant; the median correlation across facets is .47. Thus, similar results are obtained whether national stereotypes are assessed by asking about the typical culture member in general (the PPOC NCS criteria; Terracciano et al., 2005), or by averaging across assessments of national character in different age and gender groups (the present study). Because these data were obtained from entirely different sets of raters, national character stereotypes of personality appear to be generalizable.

The last three columns report correlations with the personality assessment criteria of accuracy. Significant correlations are found for 19 of the 105 correlations, and two-thirds of the correlations are positive. Results replicate across at least two of the three comparisons for five facets: Angry Hostility, Vulnerability, Tender-Mindedness, Order, and Deliberation.[5] The median facet values, however, are quite modest, ranging from .05 to .12. National character stereotypes might be said to contain a grain of truth, with true national differences accounting for perhaps 1% or 2% of the cross-national variation in perceived personality differences.

3.3. Fidelity and accuracy of stereotypes for cultures

For the profile analyses, we correlated profiles based on the 30 facets within each culture separately; this allows us to determine whether national character stereotypes are faithful and accurate for particular cultures. As in previous profile analyses (McCrae & Terracciano, 2005b; McCrae et al., 2010), both NCS means and criterion means were restandardized within the subset of cultures analyzed. Standardization ensures that profile agreement is not an artifact of normative levels of traits (see Furr, 2008); the

[5] Terracciano et al. (2005) also found evidence of accuracy for stereotypes of Vulnerability (with the observer rating criterion) and Tender-Mindedness (with the self-report criterion), but not for Angry Hostility, Order, or Deliberation.

Table 2
Associations of APPOC national stereotype profiles with PPOC national stereotype profiles and with NEO Inventory personality profiles.

Culture	Pearson *r*				*ICC*			
	NCS PPOC[a]	NEO Inventory			NCS PPOC	NEO Inventory		
		Self[b]	PPOC[c]	APPOC[d]		Self	PPOC	APPOC
Argentina	.67***	—	.19	.19	.65***	—	.19	.16
Australia	.72***	—	.48**	−.19	.65***	—	.43*	−.19
Chile	.65***	—	−.15	.19	.51**	—	−.25	.06
Croatia	.74***	−.04	−.02	.23	.72***	−.09	−.11	.11
Czech Republic	.22	.12	−.25	.48**	.21	.06	−.24	.34*
Estonia	.70***	−.01	.28	.20	.67***	−.03	.27	.16
France	.28	.34*	.33*	.25	.22	.31*	.31*	.21
Hong Kong	.40*	.31*	.64***	.43**	.20	.26	.62***	.38*
India	.55***	−.05	.04	—	.52**	−.21	−.07	—
Iran	—	—	.23	−.28	—	—	.13	−.23
Italy	.50**	.14	.27	—	.49**	.13	.23	—
Japan	.33*	−.27	.02	−.16	.25	−.29	.01	−.14
Malaysia	.42*	.08	.09	−.03	.41*	.06	.08	−.10
New Zealand	.84***	—	.63***	—	.81***	—	.63***	—
PR China	.59***	−.13	−.03	.14	.52**	−.13	−.09	.12
Peru	.06	.36*	.26	.18	.00	.19	.09	.18
Poland	.67***	.30	.49**	.20	.65***	.29	.47**	.17
Portugal	.26	−.43	−.17	−.44	.23	−.43	−.17	−.37
Russia	−.13	−.35	−.05	.14	−.20	−.36	−.06	.14
Serbia	.51**	−.05	−.04	.07	.23	−.31	−.28	−.20
Slovakia	.25	—	.00	.33*	.25	—	−.04	.07
South Korea	.23	.28	.11	.37*	.22	.20	.10	.37*
Switzerland	.71***	−.38	.03	—	.71***	−.44	−.12	—
UK	.14	—	.09	—	.12	—	−.06	—
USA	.43**	.25	.33*	−.01	.38*	.18	.24	−.05
Uganda	.50**	—	−.07	−.26	.20	—	−.25	−.40
Mdn	.50	.04	.09	.18	.38	.02	.05	.11

Note: $N = 30$ facets. Associations with NCS assess stereotype fidelity; associations with NEO Inventory assess accuracy. APPOC = Adolescent Personality Profiles of Cultures project. PPOC = Personality Profiles of Cultures project. NCS = National Character Survey.

[a] Ratings of typical culture member from Terracciano et al. (2005).
[b] Self-reports from McCrae (2002) and McCrae and Terracciano (2008).
[c] Peer ratings from McCrae and Terracciano (2005b).
[d] Peer ratings from De Fruyt et al. (2009).

* $p<.05$, one-tailed.
** $p<.01$, one-tailed.
*** $p<.001$, one-tailed.

expected value for chance agreement in these analyses is zero, and any positive correlation suggests some degree of accuracy. Further, standardization means that each facet is given equal weight in the profile. Furr (2010) recommended Pearson correlations of standardized scores as a measure of profile agreement; because they are sensitive only to the shape of the profile, they would correctly assess accuracy if the RGE attenuates the observed level of traits. In previous research we used intraclass correlations, which are more conservative because they take into account both shape and elevation of the profile. Here we report both.

Table 2 reports analyses of personality profiles for individual cultures. The first four data columns report Pearson correlations. With regard to stereotype fidelity, the profile correlations are positive in all countries except Russia;[6] 17 of 25 values (68%) are significant ($N = 30$ profile elements, $p < .05$, one-tailed), and the median correlation is .50. By contrast, the accuracy analyses in the next three columns show only 13 of 65 correlations (20%) are significant; median values ranged from .04 to .18. Replicated effects are seen only for France and Hong Kong (although Poland, which had shown evidence of an accurate autostereotype in previous research, does show a significant effect when compared to PPOC assessments). As expected, the more conservative intraclass correlations in the last four columns are consistently smaller than the Pearson correlations. Of the 25 tests of fidelity, 13 (52%) were significant, but only 9 of 65 tests of accuracy (14%) reached statistical significance. In most of these cultures, personality stereotypes contain no more than a grain of truth.

3.4. Accuracy of age- and gender-specific national stereotypes

The analyses of accuracy presented in the previous sections employed a composite stereotype score averaged across all six age-by-gender targets, but it is possible that stereotypes of some target groups may be more accurate than others. To examine that, we created a summary index of national character stereotype accuracy by correlating the profile of APPOC NCS scores with the profile of PPOC NEO Inventory means for each trait in each culture ($N = 30$ facets $\times$ 26 culture = 780 cases). For the composite stereotype score, this correlation was .12, comparable to the median facet correlation with the PPOC criteria in Table 1. Overall accuracy correlations using centered scores for each target group ranged from .00 for adolescent boys to .14 for adult women.

However, higher correlations might be found if more differentiated analyses were conducted. For this purpose, we correlated centered NCS scores for each of the six target groups with six different criteria: mean NEO Inventory scores for adolescent boys and girls from the APPOC study and college-age and adult men and women from the PPOC study. Because these are profile analyses, we used the restandardized criterion scores described in the previous section. Results are reported in Table 3. In general, the data support the hypothesis that specific stereotypes show greatest accuracy when evaluated against corresponding criteria. All 8 such matches in Table 3 are significant (Ns = 600–780 cases, $p < .05$, one-tailed),

[6] NCS Russian autostereotype data were also collected in a much larger ($N = 3695$) and more representative sample who rated Russians from their own region (Allik et al., 2009). The profile from that study correlated .49 ($p < .01$) with the PPOC Russian stereotype, so the Russian ratings in the present study appear to be anomalous.

Table 3
Profile correlations of age- and gender-specific national stereotypes with age- and gender-specific NEO Inventory criteria.

NCS target	APPOC[a] Criteria		PPOC[b] Criteria			
	Adolescent		College-age		Adult	
	Girls	Boys	Women	Men	Women	Men
Girls	**.11****	.12**	**.11*****	.08*	.04	.06
Boys	.10**	**.17*****	.04	**.12*****	−.08	−.01
Women	.01	.04	.04	.06	**.09****	.16*****
Men	−.04	.03	.00	.06	.05	**.12*****
Old Women	−.03	−.03	.04	.01	**.15*****	.20*****
Old Men	−.04	.05	−.08	−.05	.04	**.13*****

Note: Correlations in boldface are matched on age and gender. *Ns* vary because data were available from different cultures for APPOC and PPOC criteria, and because French stereotypes were available only for boys and girls. APPOC = Adolescent Personality Profiles of Cultures project. PPOC = Personality Profiles of Cultures project. NCS = National Character Survey.

[a] *Ns* = 600–630.

[b] *Ns* = 750–780.

* $p < .05$, one-tailed.

** $p < .01$, one-tailed.

*** $p < .001$, one-tailed.

with a median value of .12, whereas only 5 of 28 mismatches are significant, with a median value of .04.[7] The lowest accuracy values (−.08) are found when the stereotype of boys is compared to the assessed profile of adult women and when the stereotype of old men is compared to the assessed profile of college-age women. These data are consistent with the view that individuals are able to discern personality characteristics of specific groups—although the level of accuracy is extremely modest.

4. Discussion

All previous research using the NCS has asked for undifferentiated ratings of the typical citizen of a country or region; in this replication we specified the age and gender of the culture member. This modification yielded averaged scores that were generally comparable to those found with a global target, adding to the evidence that NCS scores yield faithful representations of shared beliefs about national character. However, consistent with most previous literature, the accuracy of these beliefs appeared to be extremely limited. Although positive associations outnumbered negative ones, the median values hovered near the summary accuracy score of .12—far lower than the corresponding summary accuracy scores of .74 for age stereotypes (Chan et al., 2012) and .67 for gender stereotypes (Löckenhoff et al., 2013). People appear to have a fairly good grasp of real age and sex differences in personality, but a largely illusory understanding of national differences.

A few traits (Angry Hostility, Vulnerability, Tender-Mindedness, Order, and Deliberation) showed significant effects in more than one analysis, and the effects for Vulnerability and Tender-Mindedness were replicated in an earlier study (Terracciano et al., 2005). It is possible that these are flukes, but researchers who wish to pursue the question of accuracy in judgments of national character might focus on these traits. Across studies, only Poland consistently appears to have a relatively accurate national character stereotype.

4.1. Collective indicators of personality

These findings seem to be at odds with data showing that stereotypes of some traits are accurate predictors of certain culture-level indicators. In particular, Heine and colleagues (2008; see also Oishi and Roth, 2009) reported that national stereotypes of Conscientiousness were associated with rapid pace of life, longevity, and *per capita* Gross Domestic Product (GDP). If we assume that these indicators reflect high collective levels of Conscientiousness, then national stereotypes appear to be accurate assessments of that trait. In contrast, these criteria were not associated with assessed Conscientiousness, and efforts to validate culture-level traits using collective behavioral indicators have had mixed success. Oishi and Roth (2009) found evidence for Agreeableness and Neuroticism, but not Conscientiousness; Mõttus, Allik, and Realo (2010) found some evidence for the criterion-related validity of Conscientiousness and its facets—although many predicted associations were not found.

However, the association between traits and outcomes is complex even at the individual level (e.g., Epstein, 1979), and likely to be much more so at the culture level (McCrae & Terracciano, 2008; Mõttus et al., 2010), where group behaviors also reflect history, government policy, religion, climate, and so on. Judged by the synchronized watches, polished shoes, and disciplined marching on a military base, one might imagine that soldiers are especially high in Conscientiousness; this, however, is not the case (e.g., Jackson, Thoemmes, Jonkmann, Lüdtke, & Trautwein, 2012). Behaviors are poor indicators of personality traits in strong situations, and culture is surely a strong situation.

Consider an instance in which stereotypes were belied by first-hand acquaintance. Asian Americans tend to be accurately stereotyped as strong academic achievers (Caplan, Choy, & Whitmore, 1989; Chao, Chiu, Chan, Mendoza-Denton, & Kwok, 2013), and one might expect that they would score high on measures of Achievement Striving and other facets of Conscientiousness. Indeed, when Anthropologist April Leininger (2002) tested a sample of Vietnamese Americans, she was initially surprised to find that as a group they scored a bit below average on measures of Conscientiousness. After several months of participant observation, however, she concluded that the scores were accurate, and she attributed the high academic achievement of her informants not to intrinsic motivation but rather to relentless pressure from family and peers to get ahead through academic pursuits in order to advance family interests. As this example illustrates, collective behaviors and group-level outcomes may not be useful criteria for assessing the accuracy of national character stereotypes because they may not reflect the operation of personality traits.

Is *per capita* wealth a good indicator of Conscientiousness? One useful test was provided by Rentfrow, Gosling, and Potter (2008), who related mean Conscientiousness levels to wealth across the 50 US states. That comparison ought to be free of any distortion by the RGE, because all Americans presumably use the same American standard of reference. Yet the observed correlation was −.19, suggesting that wealth is in fact a poor criterion of collective Conscientiousness.

McCrae et al. (2007) offered a different interpretation of the association of stereotypes of Conscientiousness with GDP: an attributional bias. Fiske, Cuddy, Glick, and Xu (2002) showed that people tend to assume that wealthy individuals are conscientious, as if their high status were clear evidence of their innate competence. In the same way, raters may presume that wealthy nations have industrious citizens. Most people know which nations are wealthy and which are poor; if knowledge of national wealth leads to presumptions of elevated national Conscientiousness, it is understandable that stereotypes of Conscientiousness are correlated with GDP. It does not, however, mean that they are accurate accounts of the personal dispositions of culture members.

Understanding the behavioral and institutional manifestations of collective personality traits in different cultural contexts is clearly an important endeavor for both personality and

[7] We conducted similar analyses at the domain level, with similar results. The median values for matched groups were .10, .07, .10, .17, and .27 for Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness facet profiles, respectively; the corresponding values for mismatched groups were .10, −.07, −.07, .08, and .13.

cross-cultural psychologists, but it is an exceptionally difficult one. We do not yet understand whether traits cause associated features of culture or vice versa (Hofstede & McCrae, 2004), and we can only speculate on how traits might interact with preexisting customs and the current socio-political situation to shape national patterns of behavior. It seems clear that there is no one-to-one correspondence between collective traits and collective behaviors that would allow us to use the latter as a gold standard criterion for validating culture-level personality measures. A more fruitful approach at this time might be the use of alternative methods that minimize RGEs (such as forced-choice measures; Heine et al., 2002) to refine our assessments of culture-level traits. We also believe that psychologists may more rapidly begin to untangle the riddles of personality and culture if they adopt the working hypothesis that current assessments of culture-level personality traits are reasonably accurate, and trace out the implications of that assumption in accounting for cultural variations in behavior and institutions. This approach might clarify which aspects of culture are and are not direct reflections of aggregate personality—and why—and may eventually lead to better assessments of the personality profiles of cultures.

4.2. Problems with national character stereotypes

If the only data by which to evaluate the accuracy of national stereotypes were correlations with assessed personality levels, critics might point to lingering doubts about the validity of personality comparisons across cultures. But in fact there are several other lines of evidence that suggest there are serious problems with national character stereotypes. In contrast to assessed personality means, national stereotypes often make no geographical sense (McCrae et al., 2007). Judging by stereotypes, Canadians are far more like Indians and Burkinabè than they are like Americans; Chinese from Hong Kong resemble Hungarians more than they resemble Chinese from the Mainland. Again, stereotypes are strongly influenced by variables, such as climatic temperature, that have no plausible relation to underlying personality traits. Stereotypes of interpersonal warmth are closely related to annual temperature ($r = .54$, $N = 49$ cultures, $p < .001$), which appears to be an effect of metaphoric thinking (McCrae et al., 2007; cf. Zhong & Leonardelli, 2008). In conjunction with contrast effects, such thinking can lead to absurd results. Southern Italy is only a few degrees warmer than Northern Italy, but stereotypes of Southern Italians portray them as over 1.5 standard deviations higher in interpersonal warmth (McCrae et al., 2007). And although Northern Italians constitute half the population of Italy, the stereotype of Northern Italians is virtually the mirror image of Italians in general ($ICC = -.72$).

The sharp contrast of regional stereotypes in Italy points to another characteristic feature of national stereotypes: They appear to exaggerate differences (Terracciano et al., 2005). Hřebíčková and Graf (2013) compared stereotypes of Austrians, Germans, Czechs, Slovaks, and Poles with personality assessments in these countries and concluded that "stereotypical beliefs exaggerate the differences between typical representatives of given countries, while their inhabitants are actually similar in most of the examined characteristics". Quantifying exaggeration is difficult when different instruments (such as the NCS and the NEO Inventories) are used, because both must be standardized in order to make them comparable. For that reason, an analysis by Realo et al. (2009) is of particular interest. They assessed both national stereotypes and self-reports of personality using the same NCS items, and compared the proportion of variance in unstandardized scores accounted for by national differences between Estonians, Latvians, Lithuanians, Poles, Belarusians, and Finns. They found that "differences in ingroup stereotype ratings between six cultural samples were six times bigger than differences in self-ratings of personality" (p. 237).

Exaggeration is probably the wrong term to apply here, because it suggests the unwarranted magnification of real differences, when in many cases there is reason to think that the perceptions of differences are completely groundless. Canadians and Americans, for example, have highly similar assessed personality profiles ($ICC = .66$), whereas their autostereotypes are diametrically opposed ($ICC = -.53$; Terracciano et al., 2005). National character stereotypes might more properly be said to exhibit unrealistically large national differences.

Finally, there is evidence that RGE can affect heterostereotypes. There is general agreement across cultures on the view that Americans are high in competence, presumably because the US is a very wealthy nation (McCrae et al., 2007). But there are also cultural variations in the degree of competence ascribed to Americans, and ratings of American competence are strongly inversely associated with the *per capita* GDP of the raters' nation, $r = -.51$, $N = 48$, $p < .001$ (Chan et al., 2011). Raters from wealthy nations judge wealth, and therefore competence, in terms of their own standards.

That finding is a reminder that national character heterostereotypes present a rich field for study, even if they prove to be as inaccurate as autostereotypes. In an ideal design, informants from a wide variety of cultures would provide ratings of each of the other cultures. Evidence to date (e.g., Boster & Maltseva, 2006; Terracciano & McCrae, 2007) suggests there would be some degree of consensus on the description of each target nation, but also some variation across cultures. That variation might be due to features of the perceivers' country, as Chan et al. (2011) showed, or to complex interactions of perceiver and target nations. For example, Argentines might have particularly negative views of the British because of the Falkland Islands conflict.

Consensual stereotypes of personality traits of different age groups and sexes have proven to be remarkably accurate. Consensual stereotypes of national character are internally consistent, generalizable across raters, and stable over time—but they show only weak traces of accuracy. New theories are needed to help explain differential accuracy in the formation of stereotypes.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Institute on Aging. Anu Realo and Jüri Allik were supported by Grants from the Estonian Ministry of Education and Science (SF0180029s08 and IUT2-13). Martina Hřebíčková and Sylvie Graf were supported by a Grant from the Czech Science Foundation (13-25656S). Robert R. McCrae and Paul T. Costa, Jr., receive royalties from the NEO Inventories.

References

Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology, 35*, 13–28.

Allik, J., Mõttus, R., Realo, A., Pullmann, H., Trifonova, A., McCrae, R. R., & Meshcheyakov, R. S., 55 Members of the Russian Character and Personality Survey. (2009). Конструирование национального характера: Свойства личности, приписываемые типичному русскому. *Cultural and Historical Psychology* (1), 2–18 (How national character is constructed: Personality traits attributed to the typical Russian).

Allport, G. W. (1979). *The nature of prejudice* (2nd ed.). New York: Perseus Books (Original work published 1954).

Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology, 44*, 61–83.

Benet-Martínez, V., & John, O. P. (1998). *Los cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*, 729–750.

Boster, J. S., & Maltseva, K. (2006). A crystal seen from each of its vertices: European views of European national characters. *Cross-Cultural Research, 40*, 47–64.

Brown, R. (2010). *Prejudice: Its social psychology*. West Sussex, UK: John Wiley & Sons.

Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American Psychologist, 22*, 817–829.

Caplan, N. S., Choy, M. H., & Whitmore, J. K. (1989). *The boat people and achievement in America: A study of family life, hard work, and cultural values*. Ann Arbor, MI: University of Michigan Press.

Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., Löckenhoff, C. E., De Bolle, M., et al. (2012). Stereotypes of age differences in personality traits: Universal and accurate? *Journal of Personality and Social Psychology, 103*, 1050–1066.

Chan, W., McCrae, R. R., Rogers, D. L., Weimer, A. A., Greenberg, D. M., & Terracciano, A. (2011). Rater wealth predicts perceptions of outgroup competence. *Journal of Research in Personality, 45*, 597–603.

Chao, M. M., Chiu, C. Y., Chan, W., Mendoza-Denton, R., & Kwok, C. (2013). The model minority as a shared reality and its implications for interracial perceptions. *Asian American Journal of Psychology, 4*, 84–92.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology, 101*, 1068–1089.

Church, A. T., & Katigbak, M. S. (2002). The Five-Factor Model in the Philippines: Investigating trait structure and levels across cultures. In R. R. McCrae & J. Allik (Eds.), *The Five-Factor Model of personality across cultures* (pp. 129–154). New York: Kluwer Academic/Plenum Publishers.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., McCrae, R. R., & Kay, G. G. (1995). Persons, places, and personality: Career assessment using the Revised NEO Personality Inventory. *Journal of Career Assessment, 3*, 123–139.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322–331.

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., & Costa, P. T. Jr.,43 Collaborators of the Adolescent Personality Profiles of Cultures Project. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment, 16*, 301–311. doi:10.1177/1073191109333760.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*, 1097–1126.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. San Diego: EdITS.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878–902.

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality, 76*, 1267–1316.

Furr, R. M. (2010). The double-entry intraclass correlation as an index of profile similarity: Meaning, limitations, and alternatives. *Journal of Personality Assessment, 92*, 1–15.

Gelade, G. (2013). Personality and place. *British Journal of Psychology, 104*, 69–82.

Guimond, S., Brunot, S., Chatard, A., Garcia, D. M., Martinot, D., Branscombe, N. R., et al. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology, 92*, 1118–1134.

Heine, S. J., & Buchtel, E. E. (2009). Personality: The universal and the culturally specific. *Annual Revuew of Psychology, 60*, 369–394.

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of Conscientiousness. *Psychological Science, 19*, 309–313.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group problem. *Journal of Personality and Social Psychology, 82*, 903–918.

Hofstede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research, 38*, 52–88.

Hřebíčková, M., & Graf, S. (2013). Accuracy of national stereotypes in central Europe: Outgroups are not better than ingroup in considering personality traits of real people. *European Journal of Personality* (doi:10.1002/per.1904).

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.

Ibrahim, F., Manaf, N. A., Kit, T. L., Tamam, E., Hilmi, K., & Darman, Z. (2010). Re-visiting Malay stereotypes: A case study among Malaysian and Indonesian Chinese students. *SEGi Review, 3*(2), 153–163.

Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtke, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military? *Psychological Science, 23*, 270–277.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system justification and the production of false consciousness. *British Journal of Social Psychology, 33*, 1–27.

Jussim, L. J. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. New York, NY: Oxford University Press.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.

Kouřilová, S., & Hřebíčková, M. (2011). Accuracy of Slovak national stereotypes: Result of judgment or intuition? *Studia Psychologica, 53*, 201–213.

Leininger, A. (2002). Vietnamese–American personality and acculturation: An exploration of relationships between personality traits and cultural goals. In R. R. McCrae & J. Allik (Eds.), *The Five-Factor Model of personality across cultures* (pp. 189–218). New York: Kluwer Academic/Plenum Publishers.

Lippmann, W. (1991). *Public opinion*. New Brunswick, NJ: Transaction Publishers (Original work published 1922).

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., et al. (2013). Gender stereotypes of personality: Universal and accurate? (submitted for publication).

Löckenhoff, C. E., De Fruyt, F., Terracciano, A., McCrae, R. R., De Bolle, M., Costa, P. T. Jr., et al. (2009). Perceptions of aging across 26 cultures and their culture-level correlates. *Psychology and Aging, 24*, 941–954.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology, 36*, 929–940.

McCrae, R. R. (2002). NEO-PI-R data from 36 cultures: Further intercultural comparisons. In R. R. McCrae & J. Allik (Eds.), *The Five-Factor Model of personality across cultures* (pp. 105–125). New York: Kluwer Academic/Plenum Publishers.

McCrae, R. R., & Costa, P. T. Jr., (2010). *NEO Inventories professional manual*. Odessa, FL: Psychological Assessment Resources.

McCrae, R. R., & Terracciano, A. (2008). The Five-Factor Model and its correlates in individuals and cultures. In D. A. van Hemert, F. J. R. Van de Vijver, & Y. H. Poortinga (Eds.), *Multilevel analyses of individuals and cultures* (pp. 249–283). Mahwah, NJ: Erlbaum.

McCrae, R. R., & Terracciano, A.78 Members of the Personality Profiles of Cultures Project. (2005a). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*, 547–561. doi:10.1037/0022-3514.88.3.547.

McCrae, R. R., & Terracciano, A.79 Members of the Personality Profiles of Cultures Project. (2005b). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*, 407–425. doi:10.1037/0022-3514.89.3.407.

McCrae, R. R., Terracciano, A., De Fruyt, F., De Bolle, M., Gelfand, M. J., & Costa, P. T. Jr.,42 Collaborators of the Adolescent Personality Profiles of Cultures Project. (2010). The validity and structure of culture-level personality scores: Data from ratings of young adolescents. *Journal of Personality, 78*, 815–838. doi:10.1111/j.1467-6494.2010.00634.x.

McCrae, R. R., Terracciano, A., Realo, A., & Allik, J. (2007). Climatic warmth and national wealth: Some culture-level determinants of national character stereotypes. *European Journal of Personality, 21*, 953–976.

McCrae, R. R., Terracciano, A., Realo, A., & Allik, J. (2008). Interpreting GLOBE societal practice scales. *Journal of Cross-Cultural Psychology, 39*, 805–810.

McCrae, R. R., Yik, M. S. M., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology, 74*, 1041–1055.

Mõttus, R., Allik, J., & Realo, A. (2010). An attempt to validate national mean scores of Conscientiousness: No necessarily paradoxical findings. *Journal of Research in Personality, 44*, 630–640.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., et al. (2012a). Comparability of self-reported Conscientiousness across 21 countries. *European Journal of Personality, 26*, 303–317. http://dx.doi.org/10.1002/per.840.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., & Johnson, W. (2012b). The effect of response style on self-reported Conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*, 1423–1436.

Oishi, S., & Roth, D. P. (2009). The role of self-reports in culture and personality research: It is too early to give up on self-reports. *Journal of Research in Personality, 43*, 107–109.

Pashler, H., & Wagenmakers, E. -J. (Eds.) (2012). *Replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science* (vol. 7(6)). (Special Section).

Peabody, D. (1985). *National characteristics*. New York: Cambridge University Press.

Perugini, M., & Richetin, J. (2007). In the land of the blind, the one-eyed man is king. *European Journal of Personality, 21*, 977–981.

Piedmont, R. L., Bain, E., McCrae, R. R., & Costa, P. T. Jr., (2002). The applicability of the Five-Factor Model in a Sub-Saharan culture: The NEO-PI-R in Shona. In R. R. McCrae & J. Allik (Eds.), *The Five-Factor Model of personality across cultures* (pp. 155–173). New York: Kluwer Academic/Plenum Publishers.

Piedmont, R. L., & Chae, J. H. (1997). Cross-cultural generalizability of the Five-Factor Model of personality: Development and validation of the NEO-PI-R for Koreans. *Journal of Cross-Cultural Psychology, 28*, 131–155.

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality, 40*, 99–120.

Realo, A., Allik, J., Lönnqvist, J.-E., Verkasalo, M., Kwiatkowska, A., Kööts, L., et al. (2009). Mechanisms of the national character stereotype: How people in six neighboring countries of Russia describe themselves and the typical Russian. *European Journal of Personality, 23*, 229–249.

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science, 3*, 339–369.

Ryan, C. S. (2002). Stereotype accuracy. *European Review of Social Psychology, 13*, 75–109.

Schmitt, D. P., Allik, J., McCrae, R. R., Benet-Martínez, V., Alcalay, L., Ault, L., et al. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*, 173–212.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cross-cultural communication style. *Journal of Cross-Cultural Psychology, 35*, 50–61.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111*, 42–61.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes businesses, economies, societies, and nations*. New York: Doubleday.

Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology, 66*, 21–36.

Terracciano, A., Abdel-Khalak, A. M., Adam, N., Adamovova, L., Ahn, C.-k., Ahn, H.-n., et al. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*, 96–100.

Terracciano, A., & McCrae, R. R. (2007). Perceptions of Americans and the Iraq invasion: Implications for understanding national character stereotypes. *Journal of Cross-Cultural Psychology, 38*, 695–710.

Zecca, G., Verardi, s., Antonietti, J.-P., Dahourou, D., Adjahouisso, M., Ah-Kion, J., et al. (2013). African cultures and the Five-Factor Model of personality: Evidence for a specific pan-African structure and profile? *Journal of Cross-Cultural Psychology, 44*, 684–700.

Zhong, C. B., & Leonardelli, G. J. (2008). Cold and lonely: Does social exclusion literally feel cold? *Psychological Science, 19*, 838–842.